A Rasch Model for Detecting Learning While Solving an Intelligence Test

Tom Verguts and Paul De Boeck
University of Leuven

A dynamic extension of the Rasch model (Verhelst & Glas, 1993, 1995) is developed from a Bayesian point of view, and it is shown how this permits application of the model in a wide variety of test settings. In particular, the method allows for an adequate modeling of learning throughout a test, determining whether learning has occurred and whether individual differences in learning rate should be assumed. An example is provided in which the model is applied to a computer-administered intelligence test. A satisfactory fit of the model was found for these data. Results indicated that learning did occur, and that there might be individual differences in learning rate. *Index terms: Bayesian statistics, dynamic Rasch model, intelligence tests, learning, Rasch model.*

Intelligence tests commonly assume that one or more latent variables (abilities) can adequately explain the responses on the test items. This logic is incorporated in the Rasch model for binary (correct/incorrect, 1/0) data. According to this model,

$$P(X_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)} , \tag{1}$$

where

$P(X_{pi} = 1)$ is the probability of a correct answer for person p on item i,
θ_p is the ability of person p, and
β_i is the difficulty of item i.

If this model does not fit the data, one solution is to assume a higher-dimensional latent space. This can be done by replacing θ_p with $\Sigma_k b_{ik}\theta_{pk}$ (e.g., McKinley & Reckase, 1983), where θ_{pk} is the ability of person p on the kth dimension, and b_{ik} is the factor loading of item i on dimension k. Other multidimensional extensions have been proposed by Embretson (1980) and Maris (1995).

Another solution is to assume that some kind of dynamic process (changing person or item parameters) occurs throughout the test. Three such processes can be distinguished (Verhelst & Glas, 1995). First, it can be assumed that some kind of intervention is introduced between blocks of items, such that θ_p increases, or equivalently, that β_i decreases (see Embretson, 1991; Fischer, 1995). Second, some kind of item interaction can be assumed, such that solving one item makes it easier to solve another item; this might be plausible, for example, if items are based on a common stimulus (e.g., Hoskens & De Boeck, 1997; Kelderman, 1984).

A third dynamic phenomenon, and the one that is assumed here, is *learning throughout the test.* When no feedback is given to the examinee, the influence of test-related learning can be questioned. However, recent literature in the field of concept learning suggests that people can nevertheless learn to solve (i.e., become better at solving) items, even if they are unsupervised (Billman & Knutson, 1996; Kersten & Billman, 1997). This is the case for most intelligence tests. Billman

Applied Psychological Measurement, Vol. 24 No. 2, June 2000, 151–162

& Knutson (1996) and Kersten & Billman (1997) described experiments in concept learning in which participants learned to discriminate concepts without instruction or information on correctly or incorrectly answered items.

Hence, it seems plausible that in an intelligence test in which a clear correct/incorrect criterion does exist, learning might occur throughout the test. This is conditional, of course, on the test items having something in common so that examinees can extrapolate from one item to another. For example, items might require a certain set of rules that are used in the test. Carpenter, Just, & Shell (1990) showed that this is the case for the Raven Progressive Matrices Test (RPMT): Only five rules are necessary to solve all the items in the test of 36 items. It seems plausible that people solving this test might improve in their ability to solve the items as the test continues.

Scheiblechner (1972) showed that learning throughout a test can occur even without direct external feedback. Following this research, some attention has been devoted to developing models expressing noncontingent learning—learning solely dependent on the number of items previously attempted (Fischer & Formann, 1982; Spada, 1976). More specifically, these authors developed versions of the linear logistic test model (LLTM) (Fischer, 1973, 1983) that can be interpreted as noncontingent learning models. However, the present research concerns contingent learning—learning dependent on the number of previous correct answers in the test. A contingent learning model was developed by Kempf (1977), which can be seen as a generalization of the Rasch model in Equation 1.

The present research is based on another dynamic generalization (Verhelst & Glas, 1993): Learning is obscured if the items are more difficult near the end of the test. For example, in the RPMT, more rules are needed per item near the end of the test. Nevertheless, it seems theoretically more satisfying to separate the one- or higher-dimensional ability that is intended to be measured from the learning aspect. Although it is impossible to erase a person's memory after every item (Holland, 1990), it is possible to statistically control the learning process that takes place during a test. The model presented here accomplishes this objective. Moreover, it can also detect whether learning occurs in the test. In this way, the methodology can both detect and account for learning.

Of course, the speed or efficiency of learning might itself be a dimension of individual differences: People might differ in the speed at which they learn and for that reason achieve different scores on intelligence tests. The model developed here cannot measure individual differences in learning—it can only detect them. Incorporating and effectively measuring these individual differences is a difficult problem, as was noted by Verhelst & Glas (1993), and cannot be solved by simply making the learning parameter person dependent.

The Dynamic Rasch Model

The extension of Equation 1 to the dynamic Rasch model (Verhelst & Glas, 1993) is

$$P(X_{pi} = 1) = \frac{\exp(\theta_p + t_{pi}\gamma - \beta_i)}{1 + \exp(\theta_p + t_{pi}\gamma - \beta_i)} \tag{2}$$

for person $p = 1, 2, \ldots, N$ and item $i = 1, 2, \ldots, I$, where t_{pi} is the number of correct answers for person p up to item $i-1$. γ is the learning parameter (LP), which scales the effect of the number of correctly answered items (t_{pi}) on the probability correct.

According to this model, examinees learn only from the items they answer correctly; other items do not influence performance. This is probably realistic in unsupervised learning, because it is often clear when an item has been answered correctly, whereas an incorrectly answered item usually contains no information for later items. However, a more general form of the model (Verhelst & Glas, 1993) allows t_{pi} to be replaced by $i - 1$, the number of items previously attempted (resulting

in a noncontingent learning model), or even $(i-1)-t_{pi}$, the number of incorrect items. The focus of the present research, however, is on the case in which the scoring variable is t.

An LLTM for Virtual Items

This model is a special case of the LLTM because it introduces *virtual items* J_{it}, which involves presenting item i to examinees who have correctly answered exactly t items, up to item $i-1$. There are $I(I+1)/2$ such items. The difficulty of such an item (denoted as ω) is

$$\omega_{it} = \beta_i + t\gamma \ , \tag{3}$$

so that a set of $I(I+1)/2$ parameters is reduced to $I+1$ in a linear fashion, as in the LLTM.

It is often possible to cluster the test items in such a way that each cluster corresponds to one solution principle or rule. For example, Carpenter et al. (1990) found five such solution principles used throughout the RPMT. Hence, the items in that test can be grouped into five clusters according to the rule involved. It is plausible that item solving benefits only from the items already solved in that cluster (termed *local learning* by Fischer & Formann, 1982). To incorporate these ideas into the model, suppose there are L clusters of items. Then, the probability of a correct answer on the ith item of cluster l is

$$P(X^l_{pi}=1) = \frac{\exp(\theta_p + t^l_{pi}\gamma^l - \beta^l_i)}{1+\exp(\theta_p + t^l_{pi}\gamma^l - \beta^l_i)} \ . \tag{4}$$

Because the LLTM has been thoroughly studied, it seems useful to treat the model as a standard LLTM. However, the number of virtual items increases rapidly (e.g., 20 items corresponds to 210 virtual items), so only small amounts of items are tractable with current computer LLTM software. Hence, the model in Equation 4 is treated as a new dynamic Rasch model.

For the first item of every item type, the model in Equation 4 reduces to Equation 1 because $t^l_{p1}=0$ for all p and l. Therefore, estimation of β^l_1 for every l will be more accurate than for $i>1$, because γ cannot influence or contaminate its estimation.

Using the LLTM for βs

The βs can be linearly restricted as

$$\beta_i = \sum_{k=1}^{K} A_{ik}\eta_k \ , \tag{5}$$

where A_{ik} are the elements of a design matrix $\mathbf{A}$. With this parameterization, the model contains two distinct LLTM aspects: (1) the restrictions at the virtual item level (i.e., for ω, see Equation 3), and (2) the βs are themselves rewritten as a function of more basic parameters η. As a result, the model becomes more stringent, but its interpretation is clearer than a model with unrestricted βs.

Estimation

Bayesian Statistics

Parameters of the model are estimated using Bayesian methods (Albert, 1992; Box & Tiao, 1973). This approach has two primary advantages: (1) posterior probability (PP) intervals (the Bayesian analogue of confidence intervals) are generated automatically, and (2) testing the model is possible without deriving the (asymptotic) distributions of the test statistics.

In Bayesian estimation, all quantities in a model are treated as random quantities (Schervisch, 1995). That is, not just the data, but also the parameters of the model are assumed to have a specified (prior) distribution. Combining these two yields a posterior distribution for the parameter vector $\boldsymbol{\xi}$:

$$p(\boldsymbol{\xi}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\xi})p(\boldsymbol{\xi}) , \tag{6}$$

which is maximized toward $\boldsymbol{\xi}$ (e.g., Swaminathan & Gifford, 1982; Tsutakawa, 1984).

Gibbs Sampling

An alternative to determining the mode of a posterior distribution is to obtain a sample of values from Equation 6. In this framework, estimating the parameters of a model consists of obtaining, for each ξ, a sample of size M from the posterior distribution. The estimated value of ξ would then be equal to

$$\hat{\xi} = \frac{1}{M}\sum_{n=1}^{M}\hat{\xi}_n , \tag{7}$$

where $\hat{\xi}_n$ is the nth sampled value from the posterior distribution. Sampling from the posterior can be done with a procedure called the Gibbs sampler (Casella & George, 1992; Gelfand & Smith, 1990; Geman & Geman, 1984; Tanner, 1996).

The Gibbs sampler requires that parameters be sampled sequentially from their full conditional distributions (Gilks, 1996; Gilks, Richardson, & Spiegelhalter, 1996). $p[\xi_k|\boldsymbol{\xi}_{(-k)}, \mathbf{x}]$ denotes the density of ξ_k, given the data ($\mathbf{x}$) and all other parameters in the model [$\boldsymbol{\xi}_{(-k)}$]. Hence, in the Gibbs sampler, first ξ_1 is sampled from $p[\xi_1|\boldsymbol{\xi}_{(-1)}, \mathbf{x}]$, then ξ_2 from $p[\xi_2|\boldsymbol{\xi}_{(-2)}, \mathbf{x}]$, and so on. If the last parameter is sampled, the process iterates at ξ_1 and the full process is repeated until convergence is reached. (For a discussion of when convergence can be assumed to be reached, see Gelman, 1996; Gelman, Carlin, Stern, & Rubin, 1995; Tanner, 1996.)

Albert (1992; Albert & Chib, 1993) showed that it is possible to estimate the Rasch model with a Gibbs sampler in which all full conditionals are normal distributions. Therefore, estimating the Rasch model consists of sampling from normal distributions. Strictly speaking, the model Albert considered is not the Rasch but the normal ogive model, which is numerically very similar to the Rasch model (Hambleton & Swaminathan, 1985). This logic can be extended to the Rasch model if the normal distribution is replaced with the logistic distribution. Following Albert, parameters are sampled from a normal distribution. To make the parameter estimates conform to a logistic model (e.g., Equation 4), they are divided by a factor $D = 1.7$. Hence, estimated parameters presented below are given by the (normal distribution) algorithm divided by $D = 1.7$.

The algorithm considered by Albert (1992) is nonstandard in that it involves the use of latent data; that is, continuous data Z_{pi} for person p and item i that are assumed to underlie the observed dichotomous data X_{pi}. Thus, Albert's algorithm becomes a combination of a data augmentation (Tanner, 1996) algorithm and a Gibbs sampler. To make the full conditionals easy to sample from, augmenting the observed data X_{pi} with the latent data Z_{pi} is necessary.

Applied to the present context, Albert's algorithm (later extended by Janssen, Tuerlinckx, Meulders & De Boeck, in press) consists of two steps:

1. Sample the latent data Z^1_{pi} from a normal distribution $N(\theta_p + t^l_{pi}\gamma^l - \beta^l_i, 1)$ truncated at 0 at the left of the normal distribution if $X^l_{pi} = 1$, and at the right of the normal distribution otherwise.
2. Sample the parameters $\boldsymbol{\xi}$ from the set of full conditional equations $p[\xi_1|\mathbf{x}, \mathbf{z}, \boldsymbol{\xi}_{(-1)}]$, ..., $p[\xi_{K_{tot}}|\mathbf{x}, \mathbf{z}, \boldsymbol{\xi}_{(-K_{tot})}]$, where $_{K_{tot}}$ denotes the total number of parameters.

These steps are iterated until convergence is reached (see below). Considering the latent data $\mathbf{z}$ makes it possible to sample from $p(\boldsymbol{\xi}|\mathbf{x}, \mathbf{z})$, which is much easier than sampling from $p(\boldsymbol{\xi}|\mathbf{x})$. More specifically, it can be shown that sampling from the full conditional equations of Step 2 reduces to sampling from a set of univariate normal distributions.

Considering that the model is an LLTM, it can be derived that the full conditional (Gibbs) equation for an arbitrary parameter ξ_k (θ, γ, β, or η) is equal to

$$p(\xi_k) = N\left[\frac{\sum_p \sum_i B_{pik}\left(Z_{pi} - \sum_{l\neq k} B_{pil}\xi_l\right) + \frac{\mu_\xi}{\sigma_\xi^2}}{\sum_p \sum_i B_{pik}^2 + \frac{1}{\sigma_\xi^2}}\ ,\ \frac{1}{\sum_p \sum_i B_{pik}^2 + \frac{1}{\sigma_\xi^2}}\right]. \tag{8}$$

In Equation 8, μ_ξ and σ_ξ^2 represent the mean and variance, respectively, of the prior distribution for ξ, which is a normal distribution. Next, B_{pik} is an element (p, i, k) of the ("extended") design matrix, which indicates the (linear) contribution of ξ_k to the (p, i) combination. Note that the items involved here are not the real items, but rather the virtual items, so the item summation is over $I(I+1)/2$ terms. For example, suppose the parameters are ordered as (θ, η, γ), and the first (virtual) item uses only the first LLTM component. Then, the vector $(B_{p11}, \ldots, B_{p1K_{tot}})$ would equal $(0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)$, with 1 and -1 appearing at the pth and the $N+1$th positions, respectively. The definition of the matrix $\mathbf{B}$ is further clarified by the observation that

$$P(X_{pi} = 1) = \frac{\exp(\Sigma_k B_{pik}\xi_k)}{1 + \exp(\Sigma_k B_{pik}\xi_k)}\ . \tag{9}$$

(For more details, see the Appendix.) Iterating the Gibbs equations inserted in the algorithm discussed above converges to sampling from the posterior distribution $p(\boldsymbol{\xi}|\mathbf{x})$ (Tanner, 1996).

A Goodness-of-Recovery Example

Baker (1998) studied the recovery of the Gibbs sampler in estimating item response theory parameters. He compared the BILOG estimation program for the two-parameter logistic model (Mislevy & Bock, 1989) to Gibbs sampling. Baker found that BILOG estimates were less biased than the Gibbs sampler for small sample sizes. In particular, at $N = 30$ and $N = 60$, the Gibbs sampler showed large bias. However, this problem was remedied for larger sample sizes ($N = 120$ or larger). Moreover, Baker's result was observed only for the estimated discrimination indices (which are not used in the present approach), not the item difficulties. Hence, Baker's data suggest that the Gibbs sampler applied to the present model will yield good results, provided N is at least 120.

To illustrate the recovery of Gibbs sampling in the present context, for $N = 300$ examinees, θs were sampled from a normal N(0, 1) distribution. Three types of items were utilized, so $L = 3$. For each item type, $I = 15$. All item parameters β_i^l were set to 0. Each item type l was assigned its own LP γ^l, with $\gamma^1 = \gamma^2 = .30$ and $\gamma^3 = .05$. After the data were generated, parameters were sampled using the algorithm described above, until convergence was reached according to the Gelman et al. (1995) $\sqrt{\hat{R}}$ criterion. Then, a sample of 200 was taken from the posterior distribution, skipping every 10 samples. Table 1 reports, for each parameter ξ, the correlation $r = r(\xi, \hat{\xi})$, where ξ denotes the true parameter value and $\hat{\xi}$ is defined as in Equation 7. For θ, for example, $r = .95$ was calculated over $N = 300$ $(\theta_p, \hat{\theta}_p)$ pairs. This correlation was undefined for the βs, because all

$\beta_i^l = 0$. The mean deviation (MD) is

$$\text{MD} = \frac{1}{C}\sum_{c=1}^{C}\left|\hat{\xi}_c - \xi_c\right| \ , \tag{10}$$

where ξ contains C parameters (e.g., $C = N = 300$ for θ, $C = 3$ for γ). As is shown in Table 1, recovery was best for γ, then β, and then θ, based on 4,500 (300×15), 300, and 45 data values, respectively.

Table 1
r and MD for
Parameter Recovery

Parameter	r	MD
θ	.95	.25
β	–	.07
γ	1.00	.01

Next, a 95% PP interval was constructed by taking the 2.5 and 97.5 percentiles, $\xi_{(2.5)}$ and $\xi_{(97.5)}$, from the 200 samples. The corresponding mean range (MR),

$$\text{MR} = \frac{1}{C}\sum_{c=1}^{C}\xi_{c,(97.5)} - \xi_{c,(2.5)} \ , \tag{11}$$

was determined, indicating how accurately the parameters were estimated. For θ, MR = 1.53; for β, MR = .46; and for γ, MR = .08.

Testing the Model

The posterior predictive check (PPC) approach was used, as advocated by Gelman & Meng (1996). Suppose the behavior of the statistic $T(\mathbf{x})$ [or, more generally, of the *discrepancy measure* $D(\mathbf{x}, \boldsymbol{\xi})$] was to be investigated. As with estimation, this approach begins with sampling a vector $\boldsymbol{\xi}$ from the posterior distribution, possibly the same as was used for estimation. From this vector, a new replicated dataset $\mathbf{x}^{\text{rep}}$ is generated, from which a replicated measure T^{rep} or D^{rep} is calculated. If this process is repeated R times,

$$\text{PPC-}p = \frac{1}{R}\sum_{r=1}^{R} I\left(D^{\text{rep,r}} \geq D^{\text{obs,r}}\right) \ , \tag{12}$$

which is a proportion if the indicator function $I(\cdot)$ takes on the value 1 when its argument is true and 0 otherwise. If this proportion is low (e.g., below .05), the model is not in accord with the data. $D^{\text{rep,r}}$ and $D^{\text{obs,r}}$ denote the values of the discrepancy measure in the replicated and observed data, respectively, with $D^{\text{rep,r}}$ a function of $\boldsymbol{\xi}$ and $\mathbf{x}^{\text{rep}}$ [formally, $f(\boldsymbol{\xi}, \mathbf{x}^{\text{rep}})$].

A global goodness-of-fit measure to be investigated for the learning model is

$$D = \sum_{l=1}^{L}\sum_{i=1}^{I_l}\sum_{t=1}^{I_l-1}\frac{\left[n_{ti+}^l - E\left(N_{ti+}^l\right)\right]^2}{E\left(N_{ti+}^l\right) + 1/2} + \frac{\left[n_{ti-}^l - E\left(N_{ti-}^l\right)\right]^2}{E\left(N_{ti-}^l\right) + 1/2} \ , \tag{13}$$

where n_{ti+} denotes the number of examinees in score group t who answered item i correctly and l denotes the item cluster. The term 1/2 is added in the denominators in order to avoid division by

zero. The expectations E(·) are a function of $\boldsymbol{\xi}$ (as drawn from the posterior distribution). Note that Equation 13 is quite similar to the Pearson χ^2 statistic.

The LPs γ^l are not person-specific, which might be a questionable assumption. However, it is not clear whether the global measure in Equation 13 is sensitive for this type of model violation. Therefore, a measure that is specifically sensitive for individual differences in LPs is

$$V = \sum_{l=1}^{L} \text{Var}\left(T^l_{end} - T^l_{begin}\right) , \tag{14}$$

where

Var(·) is the variance function (over examinees),

T^l_{begin} is the score on the first part of the items (say, items $i = 1, \ldots, I^l/2$), and

T^l_{end} is the score on the remainder of the items.

If individual differences in learning do exist, this statistic will tend to be higher than if no such differences existed. Indeed, in the former case, some values $T^l_{end} - T^l_{begin}$ will tend to be large and some small, resulting in a large variance. Therefore, it might be useful to apply the PPC approach to Equation 14 in order to test for individual differences in learning.

Analysis of an Odd-One-Out Dataset

Data

A dataset of 40 *odd-one-out* items was presented to $N = 137$ psychology students. An odd-one-out item is an item consisting of a number of elements (always 4, in this case), in which one element does not match the other three according to a certain rule. Examinees are asked to find the nonmatching element. For example, in the item

A; B; 5; F

the odd-one-out would be "5" according to the rule that every element is a letter.

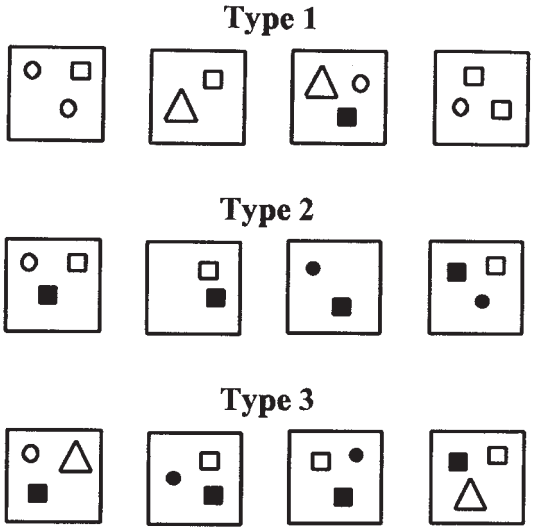
The elements of the test consisted of a combination of four kinds of geometrical figures: circles, squares, triangles, and lines. Some of these were white, some black. The test was composed of three types of items, each characterized by a certain rule that should be used to solve the item. These rules were the following for Types 1, 2, and 3, respectively: (1) the number of figures is equal to three, (2) every element contains a white figure, and (3) every element contains a black square, a white square and a third element. The number of items in each type was 13, 14, and 13, respectively. An item from each type is shown in Figure 1. The correct answers are 2, 3, and 1, respectively.

Items were computer administered in randomized order in blocks of ten items. A short break was allowed between blocks. After a response was selected, the computer informed the examinee whether that response was correct. The test items differed in the amount of ambiguity they presented. For example, the Type 1 item in Figure 1 is ambiguous because the third element might be suspected to be the odd-one-out (rather than the second) because it is the only one with a black figure. Due to the feedback provided, the participant would learn that this is an incorrect rule. Nevertheless, it seems plausible that the amount of ambiguity of an item will, to a large degree, determine the difficulty (β) of that item.

Model

The model analyzed had $L = 3$, implying that a separate LP was assigned to each item type. Moreover, it was assumed that the βs of the items could be rewritten as a function of more basic

Figure 1
Three Items of the Odd-One-Out Test



parameters (difficulties). These *basic difficulties* were assumed to be of two kinds. First, every item type was assigned its own difficulty parameter (η_1^l for type l items). Second, alternative (but incorrect) rules were assumed to partially determine the difficulty of an item, so a parameter was introduced for every alternative rule that could be used to solve the item.

The following LLTM restrictions were imposed on the βs. For Type 1, there were three factors: (1) a general "Type 1" parameter (η_1^1), (2) a parameter for the alternative rule "colored figures appear in the item," and (3) a parameter for the alternative rule "two circles in each element." The first factor was common to all Type 1 items, and indicated the difficulty of Type 1 items. The second factor was important because color was a salient but irrelevant attribute. Concerning the third factor, although some of the Type 1 items could be answered based on the rule "two circles in each element," it was an incorrect rule and always led to an incorrect answer. For Type 1 items, these three LLTM factors were the only factors involved: No item-specific parameters were introduced. For example, a Type 1 item with colored figures in which the "two circles in each element" rule was not applicable would be assigned a difficulty $\beta = \eta_1^1 + \eta_2^1$ (hence, without a η_3^1 parameter).

For Type 2 items, no plausible alternative rules could be found. Thus, a common Rasch model was assumed, resulting in 14 item parameters. Of course, it was possible to simply assign one LLTM parameter to Type 2 items, (so $\beta_i = \eta^2$ for all Type 2 items), but this restriction was too stringent because the item difficulties differed too much.

For Type 3 items, one plausible alternative rule was "two black elements in each item." Only one item, however, contained this alternative rule. Together with the basic Type 3 parameter, this resulted in two LLTM parameters for items of this type. Thus, the overall model had three LPs (because $L = 3$) and $3 + 14 + 2 = 19$ other parameters, for a total of 22 item parameters for the 40 items.

Analysis

A Gibbs sampler was run until convergence [according to the Gelman et al. (1995) $\sqrt{\hat{R}}$ criterion], from which point a sample of 1,000 vectors ($\boldsymbol{\xi}$) was produced. Each $\boldsymbol{\xi}_r (r = 1, 2, \ldots, 1{,}000)$

generated a (replicated) dataset and two discrepancy measures, $D^{\text{obs,r}}$, a function of the observed data; and $D^{\text{rep,r}}$, a function of the replicated data. D (Equation 13) and V (Equation 14) were computed, as well means for every parameter ξ (Equation 7).

Results and Discussion

PPC-p for the dynamic Rasch model without LLTM restrictions on the βs was .164 for D. This meant that the model was not rejected. Therefore, the model was restricted in the (linear) manner discussed above. As a result, PPC-$p = .000$, indicating that this restricted version of the model should be rejected. Evaluating PPC-p for every item type separately yielded PPC-$p = .151$, .029, and .000 for item Types 1, 2, and 3, respectively. Hence, Type 3 items did not conform to the restricted model. Clearly, there were more difficulty factors involved in Type 3 items than just the alternative rule "two black elements in each item."

Because Type 3 resulted in poor fit, the LLTM restrictions for Type 3 were no longer used and unrestricted item parameters (β_i) were estimated for these items. This meant that $3 + 14 + 13 + 3 = 33$ (item) parameters were estimated for 40 items. This model, for which only Type 1 items were linearly restricted, resulted in PPC-$p = .072$. Thus, this model satisfactorily fit the data and the analysis was continued using it.

Table 2 shows the mean value of γ^l for each value of l, as well as a 95% PP interval. The PP intervals indicate that learning occurred only for Type 3 items. For Types 1 and 2, γ contained 0 in its PP interval, so it was not significantly different from 0. One reason why there was no evidence for learning in Type 1 and 2 items is that these items were quite easy. The proportion of correct responses to the first Type 1 and Type 2 items were .898 and .927, respectively; consequently, there was not much improvement possible. On the other hand, the proportion correct for the first Type 3 item was .234 (and .715 for the last item), so a learning effect was possible.

Table 2
Mean and Posterior Probability (PP) Interval for γ^l, by Item Type

Item Type	Mean	95% PP Interval
1	.047	(−.052, .141)
2	−.035	(−.174, .094)
3	.535	(.313, .700)

As noted above, an interpretation is available for the basic LLTM parameters for Type 1 items. The estimated parameter values for the dynamic model are shown in Table 3. The PP intervals never contained 0 for these parameters. Thus, all effects incorporated in the model were relevant. The first parameter designated a general difficulty parameter. These items were quite easy, considering that the mean θ was .01 and the general effect (η_1^1) was −.935. The second parameter indicated the extra difficulty of the presence of color in Type 1 items. Adding color made the item more difficult: The estimated parameter value was reliably larger than zero ($\eta_2^1 = .459$). The third parameter was for items in which the rule "two circles appear" might make the item more confusing, which, indeed, was the case ($\eta_3^1 = .835$). For Type 2 and Type 3 items, there were no η parameters—only β parameters, meaning that the item difficulties could not be interpreted as stemming from basic parameters. Furthermore, no learning appeared to occur for Type 2 items.

It is interesting to compare the results of the present model with those obtained from a static LLTM (i.e., a model with $\gamma = 0$). Estimates from this model are also shown in Table 3. The general effect mean η parameter (i.e., η_1^1) was slightly smaller in this model than in the dynamic model.

Table 3
Mean and Posterior Probability (PP) Interval of η^1
for Type 1 Items From the Dynamic and Static Models

	Dynamic Model		Static Model	
η^1	Mean	95% PP Interval	Mean	95% PP Interval
1	−.935	(−1.124, −.793)	−.971	(−1.159, −.816)
2	.459	(.352, .557)	.476	(.348, .591)
3	.835	(.645, .988)	.810	(.669, .942)

This is because adding the dynamic part $t\gamma$ increased the ability parameter, which was compensated by an increase in the ηs. However, because γ was close to zero, the effect was rather small. This implies that in Type 3 items, where learning was strongest, the item parameters (β) from both models should differ most. This was indeed the case: For Type 3 items, the mean estimated β was $-.007$ in the dynamic model and $-.329$ for the static model. For Type 2 items, the estimated γ ($-.035$) was less than 0, so it would be expected that the dynamic model would result in smaller β estimates than the static model. This was the case: mean estimated βs for Type 2 items were $-.319$ and $-.262$ for the dynamic and static models, respectively. Evaluating the global goodness-of-fit statistic for the static model (without regard to item types) resulted in PPC-$p = .009$. Hence, this model fit more poorly than its dynamic counterpart.

The strong assumption was made that γ did not contain any individual differences. To investigate the validity of that assumption, Equation 14 was computed for Type 3 items, because γ was significantly different from 0 for these items. The result was PPC-$p = .414$. Hence, the assumption of no individual differences in learning was appropriate for these items. However, for items of all three types combined, PPC-$p = .029$, indicating that there appeared to be individual differences in learning rates in Type 1 and 2 items. For these item types, PPC-$p = .674$ and .000, respectively. Therefore, it can be concluded that: (1) examinees did not learn at all for Type 1 items; (2) some examinees did learn for Type 2 items, but others did not; (3) all examinees learned at approximately the same rate for Type 3 items. For Type 2 items, it seems plausible that some people learned while solving these items, whereas others performed less well when the test progressed (for example, because of fatigue), so that, on average, γ was close to 0.

Development of a model in which individual LPs are reliably estimated is still an open problem (Verhelst & Glas, 1993). However, even in the absence of a general model it seems worthwhile to have a model and methodology that can incorporate and test general learning phenomena that might be assumed to occur while responding to an intelligence test. This model and methodology can also test whether individual differences in learning do occur, although they cannot yet be incorporated as such.

Appendix: Gibbs Sampling Formulas

Z, θ, γ, and η are iteratively sampled. All parameters are assumed to have normal prior distributions [e.g., $p(\theta) \propto N(\mu_\theta, \sigma_\theta^2)$]. All prior parameters are assumed fixed. The sampling formulas are

$$p(Z_{pi}^l|\boldsymbol{\xi}, \mathbf{x}) \propto N(\theta_p + t_{pi}^l\gamma^l - \sum_k A_{ik}^l\eta_k^l, 1) \ , \tag{15}$$

truncated at the left at 0 if $X_{pi}^l = 1$, and truncated at the right at 0 if $X_{pi}^l = 0$.

Next θ, γ, and η are sampled from the distributions

$$p(\theta_p|\boldsymbol{\xi}, \mathbf{z}) \propto N\left\{ \frac{\sum_l \sum_i \left[Z_{pi}^l - \left(t_{pi}^l \gamma^l - \sum_j A_{ij}^l \eta_j^l \right) \right] + \frac{\mu_\theta}{\sigma_\theta^2}}{I + \frac{1}{\sigma_\theta^2}} , \quad \frac{1}{I + \frac{1}{\sigma_\theta^2}} \right\} , \tag{16}$$

$$p(\gamma^l|\boldsymbol{\xi}, \mathbf{z}) \propto N\left\{ \frac{\sum_p \sum_i t_{pi}^l \left[Z_{pi}^l - \left(\theta_p - \sum_j A_{ij}^l \eta_j^l \right) \right] + \frac{\mu_\gamma}{\sigma_\gamma^2}}{\sum_p \sum_i (t_{pi}^l)^2 + \frac{1}{\sigma_\gamma^2}} , \quad \frac{1}{\sum_p \sum_i (t_{pi}^l)^2 + \frac{1}{\sigma_\gamma^2}} \right\} , \tag{17}$$

and

$$p(\eta_k^l|\boldsymbol{\xi}, \mathbf{z}) \propto N\left\{ \frac{\sum_p \sum_i -A_{ik}^l \left[Z_{pi}^l - \left(\theta_p + t_{pi}^l \gamma^l - \sum_{j \neq k} A_{ij}^l \eta_j^l \right) \right] + \frac{\mu_\eta}{\sigma_\eta^2}}{\sum_p \sum_i (A_{ik}^l)^2 + \frac{1}{\sigma_\eta^2}} , \right.$$

$$\left. \frac{1}{\sum_p \sum_i (A_{ik}^l)^2 + \frac{1}{\sigma_\eta^2}} \right\} , \tag{18}$$

respectively, where the vector $\boldsymbol{\xi}$ does not contain the parameter being sampled [i.e., when sampling ξ_k, sample from $p[\xi_k|\boldsymbol{\xi}_{(-k)}, \mathbf{z})$]. Equation 8 is a general case of Equations 16–18.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251–269.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88,* 669–679.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement, 22,* 153–169.

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 458–475.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* New York: Wiley.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven Progressive Matrices Test. *Psychological Review, 97,* 404–431.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician, 46,* 19–26.

Embretson, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479–494.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–515.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3–26.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika, 60,* 459–487.

Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 6,* 397–416.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85,* 398–409.

Gelman, A. (1996). Inference and monitoring convergence. In W. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov chain monte carlo in practice* (pp. 131–143). New York: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* New York: Chapman and Hall.

Gelman, A., & Meng, X. L. (1996). Model checking and model improvement. In W. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov chain monte carlo in practice* (pp. 189–201). New York: Chapman and Hall.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721–741.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Theory and applications.* Boston MA: Kluwer-Nijhoff.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods, 2,* 261–277.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (in press). An hierarchical model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics.*

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49,* 223–245.

Kempf, W. F. (1977). Dynamic models for the measurement of traits in social behavior. In W. F. Kempf, E. B. Andersen, & B. H. Repp (Eds.), *Mathematical models for social psychology* (pp. 14–58). Bern: Huber.

Kersten, A. W., & Billman, D. (1997). Event category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 638–658.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523–547.

McKinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation, 15,* 389–390.

Mislevy, R. J., & Bock, R. D. (1989). *PC-BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. *Zeitschrift für Experimentelle und Angewandte Psychologie, 19,* 476–506.

Schervisch, M. J. (1995). *Theory of statistics.* New York: Springer-Verlag.

Spada, H. (1976). *Modelle des Denkens and Lernens.* Bern: Huber.

Swaminathan, H. & Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7,* 175–191.

Tanner, M. A. (1996). *Tools for statistical inference.* New York: Springer.

Tsutakawa, R. K. (1984). Estimation of two-parameter logistic item response curves. *Journal of Educational Statistics, 9,* 263–276.

Verhelst, N. D., & Glas, C. A. W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58,* 395–415.

Verhelst, N. D., & Glas, C. A. W. (1995). Dynamic generalizations of the Rasch model. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 181–201). New York: Springer-Verlag.

Acknowledgments

The authors thank Hans Berkhof, Eric Maris, Michel Meulders, and Francis Tuerlinckx for their helpful comments.

Author's Address

Send requests for reprints or further information to Tom Verguts, Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium. Email: Tom.Verguts@psy.kuleuven.ac.be.